# Corpus-Based Vocabulary List for Thai Language

Hathairat Ketmaneechairat [1,*] and Maleerat Maliyaem [2]

[1] College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand
[2] Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Thailand;
Email: maleerat.m@itd.kmutnb.ac.th (M.M.)
*Correspondence: hathairat.k@cit.kmutnb.ac.th (H.K.)

***Abstract*—For natural language processing, a corpus is important for training models as also for the algorithms to create the machine learning models. This paper aimed to describe the design and process in creating a corpus-based vocabulary in the Thai language that can be used as a main corpus for natural language processing research. A corpus is created under the regulation of language. By using the actual Word Usage Frequency (WUF) analyzed from a text corpus cover several types of contents. The results presented the frequency of use of several characteristics, namely the frequency of word use character usage frequency and the frequency of using bigram characters. To be used in this research and used as important information for further NLP research. Based on the findings, it was concluded that the average word length increases when the number of words in the corpus increases. It means that the correlation between word length and frequency of words is in the same direction.**

***Keywords*—corpus-based vocabulary, Thai language, frequency of words, statistical data**

## I. Introduction

The Thai language is the official national language of Thailand and is taught in elementary schools [1]. Approximately 50 million people speak Thai worldwide. It is used by about 85% of the population in Thailand, as well as by small groups of people in the United States, the United Arab Emirates, and Singapore [2]. The official Thai dictionary was published by the Royal Society of Thailand [3]. In addition to the official Thai dictionary, private dictionaries have also been created. The other important dictionaries include English-Thai dictionary of S. Sethaputra [4], and the English-Thai dictionary of Professor (Hon.) Dr. Wit Thiengburanathum [5]. Computer technology and the Internet played an important role in the creation and use of dictionaries. There has been a widespread development of electronic dictionaries both offline and online such as LEXiTRON dictionaries of the National Electronics and Computer Technology Center (NECTEC), Longdo dictionaries, Thai software dictionaries. However, there have been a few studies dealing with frequency dictionaries for Thai. The creation of a frequency dictionary of the Thai language reveals which words are most commonly used in the language and the most useful to learn Thai in the beginning. The original frequency list found on Chulalongkorn University's webpage is a version of May 2013. This version contains 5,000 words and a total frequency count of a bit above 30 million words. The main corpus categories are academic texts, non-academic, news, fiction, law, and miscellaneous, at 28.6%, 16.7%, 15.2%, 22.8%, 3.9%, and 12.6% consecutively. Words are ranked according to raw frequency [6]. Another dictionary was modified and improved by Jørgen Nilsen. This frequency dictionary is based on the "Top 5000-word lists" uploaded by Chulalongkorn University in 2013. For the Jørgen Nilsen dictionary, the proportion of most common words from 100 words, 500 words, 1,000 words, 2,000 words, 3,000 words, and 4000 words, at 47%, 71%, 81%, 91%, 95%, and 98% respectively [7].

The initial process of creating a dictionary involves word segmentation. Then many more studies focused on word segmentation have conducted with various techniques, which divided into three types, including Rules-based Word Segmentation (RBWS) [8], Dictionary-based Word Segmentation (DBWS) [9, 10], and Learning-based Word Segmentation (LBWS) [11, 12]. The performance of all algorithms are given more than 90% but there are some problems and unable to solve significant problems such as the problem with the efficiency of word segmentation, problem with compound words, problem with name entity and words with multiple spelling patterns. Moreover, a size of dictionary that contain a large number of unused words. These problems mentioned above are a significant cause of incorrect results and produce many non-word fragments in all Thai word segmentation algorithms.

In this paper describe the design and process in creating a frequency vocabulary as a corpus-based in the Thai language. The main idea is to improve the efficiency of dictionary based method by exclude excessive words from dictionary and organized words in Trie. Therefore, the word segmentation algorithm used to create the dictionary is a new algorithm that reorganized the structure of Traditional Trie to reduce the number of vocabularies and comparison tasks used in the segmentation process. By using the actual Word Usage Frequency (WUF) analyzed from a text corpus cover several types of contents, words with higher frequency are placed at the beginning of Trie that can be found and segmented more quickly. Results

Manuscript received September 5, 2022; revised October 9, 2022; accepted January 3, 2023; published April 13, 2023.

from each segmentation will also be used to update the frequency of words. Hence, the structure of Trie has improved relevant to the actual usage of each user automatically.

## II. Word Forms and Their Frequencies

### A. Structure of Entries and Frequency Data

The frequency dictionary presented in this paper includes relative word frequency information calculated from a corpus consisting of 2.2 million words. The printed version of the dictionary includes 10,000 word forms, containing 13,590 entries. The dictionary is based on newspapers, web contents, articles, reports, web boards, and chats from 53 popular websites, which related to the actual daily life usage of Thai people between the years 2016 and 2019, covering all major fields including agriculture, economics and business, society, politics, entertainment, education, health, fashion and beauty, sport, technology, and others, which are not either news or chats. Due to the size and contemporaneity of the corpus, modern written Thai is well represented. The absolute frequencies determined utilizing the corpus cannot be claimed to arrange the words of the language into a universally applicable order; a different composition of the corpus would alter the positions of almost all the words within the total scheme. Therefore, rather than giving absolute frequencies, frequency classes alone are provided in the alphabetical list. These classes are relatively stable, for, as experience shows, the frequencies of individual words in a collection of other texts compiled according to similar criteria vary by one frequency class at most. At present, spoken communication cannot be considered due to the lack of frequency data available. For the 10,000 most frequent words, the word lists are arranged in alphabetical order, each followed by its frequency class. The frequency class FC($w$) of a word "$w$" describes the frequency freq($w$) of the word in relation to the frequency $\text{freq}_{\max}$ of the most frequent word. The calculation of frequency class is in Eq. (1).

$$\mathrm{FC}(w) = \log_2 \frac{\text{freq}_{\max}}{\text{freq}(w)} \tag{1}$$

In the corpus, การ [ka:a] [work] is the most frequent word. Hence, its frequency is used as the reference frequency $\text{freq}_{\max}$. The next eight most frequent words, such as ที่ [tʰîː] [which, that], และ [lɛ́ʔ] [and], ใน [naj] [in],...,and ของ [kʰɔŋ] [of] are also belong to frequency class 0. The next seven words belong to frequency class 1. This approximately corresponds to a 50% reduction in word frequency. Similarly, ของ [kʰɔ̌ː ŋ] [of] (frequency class 0) occurs about 1,000 times more frequently than เสื้อยืด [sɯ̂ːa jɯ̂ːt] [Shirt]; (of) and เปตอง [peː tɔːŋ] [petanque] (of) (frequency class 10). This is because $2^{10}$ = 1,024. Analogously, a word belonging to frequency class 3 (e.g. มาก [mâːk] [much]) occurs about four times as often as a word in frequency class 5 (e.g. ปัจจุบัน [pàt tɕùʔ ban] [present]). More detailed information about frequencies has not been included because they are heavily dependent on the corpus in use, and thus have less significance here. Table I shows the numbers of words in different frequency classes.

TABLE I. Numbers of Words in the Different Frequency Classes

| FC | Numbers of Words |
|---|---|
| 0 | 9 |
| 1 | 7 |
| 2 | 23 |
| 3 | 54 |
| 4 | 108 |
| 5 | 256 |
| 6 | 445 |
| 7 | 706 |
| 8 | 1,034 |
| 9 | 1,427 |
| 10 | 1,763 |
| 11 | 2,204 |
| 12 | 2,510 |
| 13 | 3,014 |

### B. Data in Electronic Form

The electronic form is data about the 13,590 most frequently occurring word forms in Thai, as determined in the corpora analysis. Details about the frequency classes to which the word forms belong are also given. The data is available in two formats: as an e-book, i.e., as a PDF file in a format similar to this document and so that the data can be put to further use easily, the list of the 13,590 most frequent words will be made available as a plain text file which ordered in each of the following ways: arranged according to frequency and arranged in alphabetical order.

### C. Conception of a Dictionary of Word Forms

The frequency dictionary shows the words as they occur in the corpus. Different from many European languages, Thai words have only base forms without inflected forms caused by tense or capital letters. For this reason, the words in this dictionary are shown as they appear in the corpus without being normalized. Thai grammar relies mostly on the sequence of syllables, and sentence structure is continued without specific word boundary delimiters. Blanks are used as an arbitrary written style. A word can be composed of one or two or even more syllables. The Thai sentence structure or word order starts from left to right which is; Subject + Predicate. The Subject is the part to represent the actor, which may be a noun, pronoun, noun phrase, or sentence. In some cases, modifiers are added to describe properties or details of the Subject. The Predicate is a verb or verb phrase used to show the manner of acting of the Subject. Similar to Subject, Predicate may consist of a modifier. Moreover, some verb requires an Object to complete the sentence's meaning. Same as the Subject, an Object is a noun, pronoun, noun phrase, or sentence, which sometimes has a modifier [13]. Examples of Thai sentences are shown as follows:

In this dictionary, words are first segmented based on analyzing each sentence in the corpora before they are counted. Thai word segmentation is a very difficult task, because of the complicated rules of words and sentences forming. Moreover, all Thai words in multiple sentences are written continuously without any spaces or delimiters,

which are also found in many other non-segmentation languages such as Lao, Chinese, and Burmese. For example, the sentence สัปดาห์หน้าคณะรัฐมนตรีจะเดินทางไปประชุม (Next week, the cabinet will travel to a meeting) will be segmented into words as follow:

| สัปดาห์ | หน้า | คณะ | รัฐมนตรี | จะ | เดินทาง | ไป | ประชุม |
|---|---|---|---|---|---|---|---|
| sàp da: | nâ: | kʰá náʔ | rát mon tri: | tɕàʔ | dɤ:n tʰa:ŋ | paj | pràʔ tɕʰum |
| week | next | group | cabinet | will | travel | to | meet |

### *D. On the Definition of a Word*

The Word is the smallest unit of a language that has meaning. Each word in the Thai language may consist of one or more syllables, which are formed by three types of alphabets: Consonants, Vowels, and Tonal Marks.

*1) Thai consonants*

Thai Consonants consist of 44 alphabets, which can be divided into six groups based on the organ in the mouth that causes the sound of the alphabet:

Group 1: กัณฐชะ [kan tʰa tɕʰá] is the consonants pronounce from the base of the throat, which are: ก [k] ข [kʰ] ฃ [kʰ] ค [kʰ] ฅ [kʰ] ฆ [kʰ], and ง [ŋ].

Group 2: ตาลุชะ [ta: lu tɕʰá] is the consonants pronounce from the base of the palate, which are: จ [tɕ] ฉ [tɕʰ] ช [tɕʰ] ซ [s] ฌ [tɕʰ], and ญ [j].

Group 3: มุทธชะ [mut̄tʰa tɕʰá] is the consonants pronounce from the base of the gum with tongue, which are: ฎ [d] ฏ [t] ฐ [tʰ] ฑ [tʰ] ฒ [tʰ] ฌ [tʰ], and ณ [n].

Group 4: ทันตะชะ tʰa[n ta tɕʰá] is the consonants pronounce from the base of the teeth with tongue, which are: ด [d] ต [t] ถ [tʰ] ท [tʰ] ธ [tʰ], and น [n].

Group 5: โอฐชะ [ʔo:t̄tʰa tɕʰá] is the consonants pronounce from the base of the lip, which are: บ [b] ป [p] ผ [pʰ] ฝ [f] พ [pʰ] ฟ [f] ภ [pʰ], and ม [m].

Group 6: อวรรค [ʔà wak̄] is the consonants pronounce from the base other organs, which are: ย [j] ร [r] ล [l] ว [w] ศ [s] ษ [s] ส [s] ห [h] ฬ [l] อ [ʔ], and ฮ [h].

Currently, two consonants, including ฃ and ฅ, have been canceled. Therefore, the remainder of 42 consonants is used as initial consonants.

For final consonants, only 35 Thai consonants are used and divided into eight categories according to their pronunciation as follows:

1. The final consonants, which are pronounced as [k̄], are ก ข ค and ฆ.
2. The final consonants, which are pronounced as [t̄], are ค จ ช ซ ฎ ฏ ฐ ฑ ฒ ฌ ด ถ ท ธ ศ ส and ย.
3. The final consonants, which are pronounced as [p̄], are บ ป พ ฟ and ภ.
4. The final consonant, which is pronounced as [ŋ], is ง.
5. The final consonants, which are pronounced as [n], are น ญ ณ ร ล and ฬ.
6. The final consonant, which is pronounced as [m], is ม.
7. The final consonant, which is pronounced as [j], is ย.
8. The final consonant, which is pronounced as [w], is ว.

Nine consonants which cannot be a final consonant are ฃ ฅ ฉ ฌ ผ ฝ ห อ and ฮ.

*2) Thai vowels*

Thai vowels consist of 44 alphabets including 21 characters used solitary or combined as a diphthong to be 32 Thai vowels that are divided into two types: Short-sound vowels and long-sound vowels. Each vowel and its pronunciation are shown in Table II.

| Sentence | | Words | | | | Meaning in English |
|---|---|---|---|---|---|---|
| ฉันเดิน | → | ฉัน | เดิน | | → | I walk |
| | | [tɕʰăn] | [dɤ:n] | | | |
| | | I | walk | | | |
| นกกินหนอน | → | นก | กิน | หนอน | → | A bird eats a worm |
| | | [nók] | [kin] | [nŏ:n] | | |
| | | Bird | eat | worm | | |

TABLE II. THAI VOWELS

| Short-Sound-Vowel | Phonetic Annotation |
|---|---|
| ะ, ั | [à] |
| ิ | [ì] |
| ึ | [ù] |
| ุ | [ù] |
| เ◌ะ, เ็ | [è] |
| แ◌ะ, แ็ | [ɛ̀] |
| โ◌ะ | [ò] |
| เ◌าะ, เื่อ | [ɔ̀] |
| เ◌อะ | [ɤ̀ʔ] |
| เ◌ียะ | [ìàʔ] |
| เ◌ือะ | [ɯàʔ] |
| ◌ัวะ | [uàʔ] |
| ฤ | [rɯ́] |
| ฦ | [lɯ́] |
| ◌ำ | [am] |
| ไ◌ | [ai] |
| ใ◌ | [ai] |
| เ◌า | [au] |

| Long-Sound-Vowel | Phonetic Annotation |
|---|---|
| า | [a:] |
| ี | [i:] |
| ื | [ɯ:] |
| ู | [u:] |
| เ | [e:] |
| แ | [ɛ:] |
| โ | [o:] |
| อ, ื | [ɔ:] |
| เ◌อ | [ɤ:] |
| เีย | [ia] |
| เือ | [ia] |
| ัว | [ɯa] |
| ฤๅ | [ua] |
| ฦๅ | [rɯ:] |

*3) Tonal marks*

Tonal Marks consist of four alphabets to indicate Tonal marks. Thai words have five different levels of tone. To modify the sound of a word to a different level, four tonal

marks, as shown in Table III, are used to combine with consonants and vowels.

TABLE III. TONAL MARKS

| Tonal mark | Name of Tonal mark |
|---|---|
| ◌่ | เอก [ʔàːk] |
| ◌้ | โท [tʰoː] |
| ◌๊ | ตรี [triː] |
| ◌๋ | จัตวา [tɕàt taː waː] |

Thai words are divided into two types, which are Base-word and Compound-word. A Base-word is the smallest unit of a language that is meaningful and cannot be separated into two or more smaller words. Each Base-word may have only one syllable or many syllables. Some examples of Basic words are shown in Table IV.

TABLE IV. EXAMPLES OF BASE WORDS

| Amount of syllables | Word | Phonetic annotation | Meaning in English |
|---|---|---|---|
| 1 | กิน | [kin] | Eat |
| | เดิน | [dɤːn] | Walk |
| 2 | สะอาด | [sàʔ ʔàːt] | Clean |
| | สหาย | [sà hăːj] | Friend |
| More than 2 | สวัสดี | [sà wàt diː] | Hello |
| | อนาคต | [ʔà naː kʰót] | Future |

Compound-word is a word formed by two or more Base-words to create a new word in various combination types, as shown in Table V.

TABLE V. EXAMPLES OF COMPOUND-WORDS

| Combination | First word | | Second word | | Compound-word |
|---|---|---|---|---|---|
| Noun+ Noun | คน [kʰon] [Human] | + | สวน [sŭːan] [Garden] | → | คนสวน [kʰon sŭːan] [Gardener] |
| Noun+ Verb | ห้อง[hɔ̂ːŋ] [Room] | + | นอน [nɔːn] [Sleep] | → | ห้องนอน [hɔ̂ːŋ nɔːn] [Bed room] |
| Noun + Modifier | ข้าว[kʰâːw] [Rice] | + | สวย [sŭaj] [Beautiful] | → | ข้าวสวย [kʰâːw sŭaj] [Steamed rice] |
| Verb + Verb | เดิน[dɤːn] [Walk] | + | ทาง [tʰaːŋ] [Way] | → | เดินทาง [dɤːn tʰaːŋ] [Travel] |
| Verb+ Noun | ลง [loŋ] [Down] | + | โทษ [tʰôːt] [Punishment] | → | ลงโทษ [loŋ tʰôːt] [Punish] |
| Verb + Modifier | กิน [kin] [Eat] | + | แหลก [lɛ̀ːk] [Crushed] | → | กินแหลก [kin lɛ̀ːk] [Eat Everything] |
| Modifier+ Modifier Modifier | ดี [diː] [Good] | + | งาม [ŋaːm] [Beautiful] | → | ดีงาม [diː ŋaːm] [diː ŋaːm] [Very good] |

Thai compound words are divided into three cases as follows:

Case 1: The compound words which have the same meaning as their original base words.

Case 2: The compound words have a different meaning from their original base words.

Case 3: The compound words that are formed by a keyword, such as การ [kaːn], ความ [kʰwaːm], นัก [nák], combining with other basic words to change the type of basic words, such as the word การ [kaːn] + a verb and turn the verb into a noun, which is similar to the addition of "ing" to English verbs to become nouns.

In this dictionary, compound words are mainly applied in Case 2. In Case 1, only compound words which are generally used by Thai people accept while the others which are used by a specific group of persons will be segmented into base words. For Case 3, the keywords are separated from the base words.

## III. THE PROPOSED NEW ALGORITHM

The proposed technique, Thai Language Segmentation by Automatic Ranking Trie or TLS-ART, employs actual usage frequency to reduce the dictionary size and number of matching tasks for splitting and identifying words in sentences. There are three steps in the proposed technique as shown in Fig. 1.

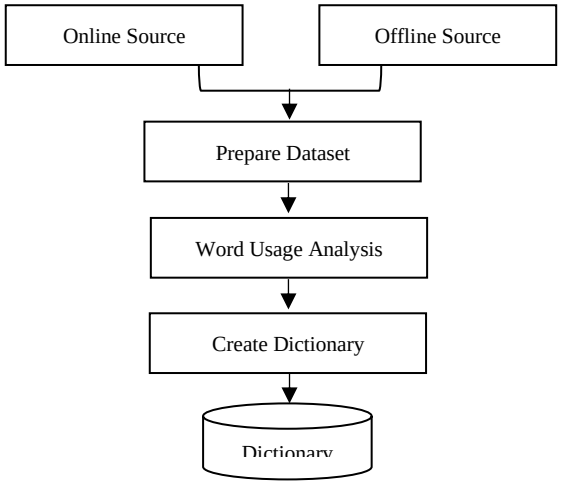


Figure 1. The proposed techniques steps.

### A. Dataset Preperation

The words and their frequencies were determined based on different sources. The dictionary is based on the following sources:

Newspapers, web contents, articles, reports, Web boards, and chats from 53 popular websites, related to the actual daily life usage of Thai people within the years 2016-2019.

Cover all major fields including agriculture, economics and business, society, politics, entertainment, education, health, fashion and beauty, sport, technology, and others, which are not either news or chats.

There are 3,586 files, shown in Table VI in detail.

Therefore, about 13,590 entries are a reasonable result included in the dictionary.

TABLE VI. NUMBER OF SAMPLE TEXT FILES IN EACH CATEGORY OF DATA SET

| Category | Number of files |
|---|---|
| Agriculture | 325 |
| Economics and business | 326 |
| Society | 328 |
| Politics | 325 |
| Entertainment | 323 |
| Education | 325 |
| Health | 327 |
| Fashion and beauty | 326 |
| Sport | 324 |
| Technology | 326 |
| Others | 331 |
| Total | 3586 |

Several data processing steps are necessary to gain a usable data set for the compilation of a frequency dictionary [14]. From this Table VI, the datasets used in this paper are 3,586 text files collected from various sources, including both online and offline consistent with real-life usage. In the case of online sources, text messages in various types of content were collected from popular websites, including news, documentaries, Wikipedia, chat rooms, government agencies, and business organizations. For offline sources, articles, journals, books, official letters, publications, reports, etc., were purposed as data sources for text message collection. Then, all collected messages were cleaned by a preprocessing process to eliminate all control codes (i.e., character style, display specification, etc.), as well as menu options, which are excessive redundant texts. Subsequently, each message was saved as a text file, based on its content type, into the Corpus.

### B. Word Usage Analysis

The next process is word segmentation. In this dictionary, the Thai Language Segmentation by Automatic Ranking Trie Model (TLS-ART) is used to create a frequency dictionary. The segmentation process of TLS-ART is divided into two steps. In the first step, a text message will be segmented into base words by parsing with the Automatic Ranking Trie. Then, in the second step, all base words are analyzed and counted for the appearance of each word to provide Word Usage Frequency (WUF). WUF analysis was conducted based on real usage data in daily life.

### C. Create Dictionary

This process saves words and WUF of each word to dictionary and creates Trie by placing ordered words based on usage frequency from high to low. Finally, the outputs are summarized to present the usage frequency in many characteristics, including Words Usage Frequency, Characters Usage Frequency, and Character-bigram Usage Frequency.

## IV. STATISTICAL DATA ON THE WORD LIST

### A. Comparison Criteria

The linguo-statistical data obtained from the analysis of a word list can shed light on various aspects of a language [15]. This dictionary series not only shows strong regularities for a particular language but provides data for different languages allowing statistical comparisons between them. The comparison criteria include the following:

- The alphabet and its letter frequencies,
- Word lengths,
- Word structure, e.g., the relationship between vowels and consonants; the number and length of syllables,
- Vocabulary range measured by text coverage,

Dependency of features, for example, frequency (Zipf's law) and word length, on the rank in the word list.

In many cases, the data allows us to infer characteristic parameters that are useful, for example, for lingo typological investigations. The parameters important for lingo statistical comparisons are shown in a Table VII. For the statistical analysis, the use of word forms is of utmost importance because the particular word forms usually differ in their length and number of syllables etc. With regard to the number of features, such as the frequency of letters or the average number of syllables, the results of measurements depend on whether each word is counted only once or whether the words are considered according to their frequencies. This is called analysis without or with multiplicity. If it was deemed reasonable, both analyses were performed. The results without multiplicity are presented in the left column or left picture, while those including multiplicity are shown on the right. The tables include data up to 2.2 million words because these are available and because the average data is still meaningful for words with low frequency.

### B. Character Statistics

The counting of characters is, in general, one of the easiest statistical analyses to carry out. The analysis result of each character, including consonants, vowels, tonal and special characters shown in the percentage of frequency usage as the detail in Table VII.

TABLE VII. CHARACTER FREQUENCIES IN THE WORD LIST AND IN THE TEXT

| Character frequency in %. | | |
|---|---|---|
| | Without multiplicity of words | With multiplicity of words |
| Special character | 0.50 | 0.40 |
| ก [K] | 4.31 | 4.53 |
| ข [$K^h$] | 0.74 | 1.08 |
| ฃ [$K^h$] | 0.00 | 0.00 |
| ค [$K^h$] | 1.73 | 1.71 |
| ฅ [$K^h$] | 0.00 | 0.00 |
| ฆ [$K^h$] | 0.04 | 0.02 |
| ง [ŋ] | 3.29 | 3.61 |
| จ [Tɕ] | 1.34 | 1.51 |
| ฉ [$Tɕ^h$] | 0.16 | 0.07 |
| ช [$Tɕ^h$] | 1.33 | 1.06 |
| ซ [S] | 0.67 | 0.32 |
| ฌ [$Tɕ^h$] | 0.00 | 0.00 |
| ญ [J] | 0.36 | 0.23 |
| ฎ [D] | 0.04 | 0.03 |
| ฏ [T] | 0.06 | 0.03 |
| ฐ [$T^h$] | 0.13 | 0.11 |
| ฑ [$T^h$] | 0.06 | 0.02 |
| ฒ [$T^h$] | 0.06 | 0.05 |

| Character frequency in %. | | |
|---|---|---|
| | Without multiplicity of words | With multiplicity of words |
| ง [N] | 0.44 | 0.35 |
| ด [D] | 2.51 | 2.53 |
| ต [T] | 2.53 | 2.00 |
| ถ [Tʰ] | 0.41 | 0.44 |
| ท [Tʰ] | 1.90 | 2.31 |
| ธ [Tʰ] | 0.56 | 0.33 |
| น [N] | 5.67 | 5.79 |
| บ [B] | 2.06 | 1.86 |
| ป [P] | 1.64 | 1.77 |
| ผ [Pʰ] | 0.38 | 0.53 |
| ฝ [F] | 0.12 | 0.06 |
| พ [Pʰ] | 1.73 | 1.27 |
| ฟ [F] | 0.40 | 0.14 |
| ภ [Pʰ] | 0.42 | 0.28 |
| ม [M] | 3.35 | 3.60 |
| ย [J] | 3.01 | 2.84 |
| ร [R] | 5.86 | 5.40 |
| ฤ [Rɯ] | 0.09 | 0.04 |
| ล [L] | 3.02 | 2.59 |
| ฦ [Lɯ] | 0.00 | 0.00 |
| ว [W] | 2.93 | 2.78 |
| ศ [S] | 0.60 | 0.37 |
| ษ [S] | 0.31 | 0.29 |
| ส [S] | 2.77 | 2.00 |
| ห [H] | 1.75 | 1.88 |
| ฬ [L] | 0.02 | 0.02 |
| อ [ʔ] | 4.57 | 4.24 |
| ฮ [H] | 0.15 | 0.04 |
| ะ [À] | 1.54 | 2.02 |
| ั [À] | 3.49 | 3.16 |
| า [A:] | 6.23 | 7.16 |
| ◌ิ [Ì] | 3.30 | 1.96 |
| ◌ี [I:] | 2.23 | 2.73 |
| ◌ึ [ɯ̀] | 0.33 | 0.49 |
| ◌ึ [ɯ:] | 0.80 | 1.10 |
| ◌ุ [Ù] | 1.70 | 1.01 |
| ◌ู [U:] | 0.93 | 0.88 |
| เ [E:] | 3.73 | 3.75 |
| แ [Ɛ:] | 1.13 | 1.51 |
| โ[O:] | 0.95 | 0.67 |
| ใ [AI] | 0.35 | 1.17 |
| ไ [AI] | 0.60 | 1.31 |
| ◌ำ [AM] | 0.65 | 0.82 |
| า [A:] | 0.00 | 0.00 |
| ◌ื [ɔ:] | 0.44 | 0.85 |
| ์ [KA:-RAN] | 1.89 | 0.65 |
| ่ [TɛÙT̄] | 0.00 | 0.00 |
| ่ [ʔÀ:K̀] | 2.78 | 4.45 |
| ้ [TʰO:] | 2.66 | 3.73 |
| ็ [TRI:] | 0.18 | 0.03 |
| ่ [TɛÀT̀ TA: WA:] | 0.06 | 0.02 |

## C. Vowels and Consonants

Languages often differ regarding their proportions of vowels and consonants as well as in terms of their sequence. The co-occurrence analysis of consonants and vowels was performed in four cases:

- Initial-consonant:Vowel
- Initial-cluster-consonant:Vowel
- Vowel:Final-consonant
- Vowel:Final-cluster-consonant

The results are showed in Tables VIII–XI.

TABLE VIII. TOP TEN OF THE MOST USAGE BIGRAMS OF INITIAL CONSONANT: VOWEL

| Initial consonant | Vowel | Percent of usage |
|---|---|---|
| ส [s] | (reduction of ะ [à]) | 1.46 |
| ร [r] | า [a:] | 1.13 |
| ก [k] | า [a:] | 1.00 |
| น [n] | า [a:] | 0.92 |
| ร [r] | ั [à] | 0.82 |
| ม [m] | า [a:] | 0.81 |
| ว [w] | ิ [i] | 0.79 |
| ส [s] | า [a:] | 0.71 |
| ย [j] | า [a:] | 0.69 |
| ต [t] | า [a:] | 0.66 |

TABLE IX. TOP TEN OF THE MOST BIGRAMS OF INITIAL CLUSTER-CONSONANT: VOWEL

| Initial consonant | Vowel | Percent of usage |
|---|---|---|
| ปร [pr] | ะ [à] | 0.80 |
| กร [kr] | ะ [à] | 0.57 |
| หน [nɔ̌] | า [a:] | 0.37 |
| พร [pʰr] | ะ [à] | 0.37 |
| หล [lɔ̌] | (reduction of ไ [ò]) | 0.19 |
| หล [lɔ̌] | ั [à] | 0.19 |
| หม [mɔ̌] | า [a:] | 0.17 |
| กล [kl] | า [a:] | 0.15 |
| ศร [sr] | ี [i:] | 0.14 |
| ปร [pr] | า [a:] | 0.14 |

TABLE X. TOP TEN OF THE MOST BIGRAMS OF VOWEL: INITIAL CONSONANT

| Vowel | Final Consonant | Percent of usage |
|---|---|---|
| ั [à] | น [n] | 3.76 |
| า [a:] | ย [j] | 2.70 |
| ั [à] | ง [ŋ] | 2.64 |
| อ [ɔ:] | ง [ŋ] | 2.27 |
| (reduction of ไ [ò]) | ง [ŋ] | 2.22 |
| า [a:] | น [n] | 2.18 |
| า [a:] | ง [ŋ] | 2.18 |
| ั [à] | ก [k] | 2.09 |
| (reduction of ไ [ò]) | ม [m] | 2.01 |
| ิ [i] | น [n] | 1.91 |

TABLE XI. TOP TEN OF THE MOST USED BIGRAMS OF VOWELS: INITIAL CLUSTER-CONSONANT

| Vowel | Final Consonant | Percent of usage |
|---|---|---|
| ั [à] | ตร [t̀] | 0.14 |
| ิ [i] | ตร [t̀] | 0.11 |
| ั [à] | กร [k̀] | 0.10 |
| ุ [ù] | ตร [t̀] | 0.08 |
| า [a:] | ตร [t̀] | 0.06 |
| เ [e:] | ตร [t̀] | 0.05 |
| (reduction of ไ [ò]) | หม [m] | 0.05 |
| ุ [ù] | ทร [t̀] | 0.04 |
| ั [à] | คร [k̀] | 0.04 |
| ู [u:] | ตร [t̀] | 0.04 |

### D. Word Length

The analysis of word length generally starts with counting the words of different lengths. If we count words with multiplicity, the number of short words will, of course, be larger, as these are the most frequent. The results of word length analysis in the term of syllable and number of characters are shown in detail in Tables XII and XIII.

TABLE XII. WORD LENGTH ANALYSIS RESULT (SYLLABLE COUNT)

| Number of Syllables | Total count without multiplicity | Total count with multiplicity |
|---|---|---|
| 1 | 4843 | 1,575,219 |
| 2 | 9457 | 1,101,444 |
| 3 | 3904 | 382,839 |
| 4 | 1383 | 128,696 |
| 5 | 329 | 29,960 |
| 6 | 103 | 22,848 |
| 7 | 20 | 4683 |
| 8 | 13 | 904 |
| 9 | 3 | 2142 |
| 10 | 3 | 1490 |
| 11 | 0 | |

TABLE XIII. WORD LENGTH ANALYSIS RESULT (CHARACTER COUNT)

| Number of Syllables | Total count without multiplicity | Total count with multiplicity |
|---|---|---|
| 1 | 1 | 973 |
| 2 | 444 | 606,624 |
| 3 | 2311 | 2,363,682 |
| 4 | 2682 | 1,783,252 |
| 5 | 2713 | 1,240,635 |
| 6 | 3437 | 1,220,448 |
| 7 | 2985 | 999,054 |
| 8 | 2048 | 551,600 |
| 9 | 1379 | 407,142 |
| 10 | 843 | 174,070 |
| 11 | 485 | 150,491 |
| 12 | 293 | 57,624 |
| 13 | 169 | 23,296 |
| 14 | 130 | 21,994 |
| 15 | 74 | 10,275 |
| 16 | 33 | 6944 |
| 17 | 10 | 3978 |
| 18 | 8 | 900 |
| 19 | 8 | 1539 |
| 20 | 1 | 680 |
| 21 | 3 | 273 |
| 22 | 1 | 176 |
| 23 | 0 | 0 |

### E. Zipf's Law

The rank of a word is defined as its position in a list of words that have been sorted according to frequency. Rank 1 is given to the most frequent word, rank 2 to the second most frequent, etc. The rank-frequency law by G.K. Zipf establishes the following relationship between the rank and frequency of a word: For a sufficiently large corpus, the product of rank and frequency is almost constant for all words. Fig. 2 shows rank and frequency in log-log coordinates.

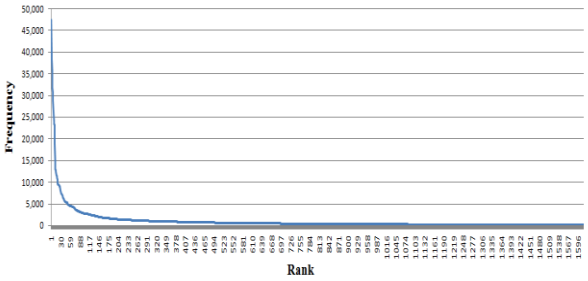


Figure 2. Zipf's law.

### F. Text Coverage of the Top-N Most Frequent Words

Text coverage indicates the proportion of a text (measured in words) that consists exclusively of the given words. In other words, if we mark a set of words, the text coverage would give the proportion of these words to the total number of words in the text. Here, we examine the text coverage of the top-N most frequent words. Fig. 3 shows the almost linear growth of text coverage with a logarithmic scaling of the N-axis for small N. This is an immediate consequence of Zipf's law.

| N | Coverage |
|---|---|
| 10 | 15.62 |
| 100 | 42.12 |
| 1000 | 78.06 |
| 5000 | 96.33 |
| 10000 | 99.36 |
| 13590 | 100.00 |

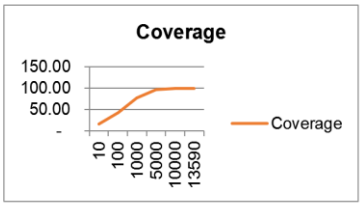


Figure 3. Text coverage of the top N most frequency words (in percentage).

### G. Growth in Word Length

The examination of word length can reveal a number of interesting results, quite apart from the fact that extremely long words have entertainment value in themselves. Table XIV provides the longest words within the most frequent N words. Using logarithmic scaling for the frequencies, the increase in word length appears almost linear. The slope of the line of best fit in the frequency range.

TABLE XIV. LONGEST WORDS WITHIN THE MOST FREQUENT N WORDS

| N | Longest word | Length |
|---|---|---|
| 10 | เป็น [pen] | 4 |
| 100 | สามารถ [sǎ: mâ:t] | 6 |
| 1000 | พระเจ้าอยู่หัว [pʰráʔ teâw jù: hŭ:a] | 14 |
| 5000 | สิริวัฒวรีนารีรัตน์ [siʔ ríʔ wan ná wá ri: na: ri: rát] | 20 |
| 10,000 | บัณฑิตพัฒนบริหารศาสตร์ [ban tʰít pʰát nóp ríʔ hă:n sà:t] | 22 |
| 13,590 | บัณฑิตพัฒนบริหารศาสตร์ [ban tʰít pʰát nóp ríʔ hă:n sà:t] | 22 |

TABLE XV. AVERAGE WORD LENGTH IN VARIOUS FREQUENCY RANGES

| N | Average Word length |
|---|---|
| 10 | 1.00 |
| 100 | 1.12 |
| 1000 | 1.54 |
| 5000 | 1.82 |
| 10,000 | 1.97 |
| 13,590 | 2.07 |

The maximum word length increases with N: As the more frequent words tend to be shorter in length, the rarer words show a tendency to become longer. To examine this connection in more detail we need to look at the average word lengths of the most frequent N words instead of just taking into account the longest word itself (see Table XV).

### H. Number of Syllables

In Thai, syllable segmentation was conducted to improve the rules using the consonant and vowel characteristics according to Thai spelling principles [9]. Words consist of one, two or more syllables. The focus of this section is the average number of written syllables per word. As word length increases for rare words, the average number of syllables also increases (see Table XVI).

TABLE XVI. AVERAGE WORD LENGTH IN VARIOUS FREQUENCY RANGES

| N | The average number of syllables per word | |
|---|---|---|
| | without multiplicity | with multiplicity |
| 10 | 1.00 | 1.00 |
| 100 | 1.12 | 1.06 |
| 1,000 | 1.54 | 1.27 |
| 5,000 | 1.82 | 1.38 |
| 10,000 | 1.97 | 1.40 |
| 13,590 | 2.07 | 1.40 |

### I. Summary of Statistical Data

Table XVII provides a summary of the most important lingo specific data. The presentation can serve as a basis for a comparison between different languages.

TABLE XVII. SUMMARY OF STATISTICAL DATA

| Parameter | Value | Section |
|---|---|---|
| Range of the characters, including consonants, vowels, tonal and special characters | 71 | II |
| Percent of the most usage bigrams of Initial consonants: Vowel | 1.46% | IV |
| Percent of the most usage bigrams of Initial cluster consonant: Vowel | 0.80% | IV |
| Percent of the most usage bigrams of Vowel: Initial consonant | 3.76% | IV |
| Percent of the most usage bigrams of Vowel: Initial cluster-consonant | 0.14% | IV |
| Text coverage of most frequent 10 words | 15.62 % | IV |
| Text coverage of most frequent 100 words | 42.12 % | IV |
| Text coverage of most frequent 1000 words | 78.06 % | IV |
| Average word length of the most frequent 10,000 words | 1.97% | IV |
| The average number of syllables of the most frequent 10,000 words (with multiplicity) | 1.40% | IV |
| The average number of syllables of the most frequent 10,000 words (without multiplicity) | 1.97% | IV |

## V. CONCLUSION

A corpus is an important part to create an effective model in NLP. This paper proposed a corpus-based vocabulary created based on regulation and statistics. This work started with data collection and data preprocessing. A statistic in terms of frequency of words is used to classify a type of parameter. Average word length and Average Number of syllables are considered to measure. The Longest words within the most frequent N words are also investigated and measured. The result of this work is a corpus-based vocabulary created and can be used for NLP in the future. In this case, Machine Learning seems to be the most suitable algorithm that can learn and extend the scope of utilization to other areas, such as language translation, the Question-answering system, and the interface between humans and computers in various ways.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

Hathairat Ketmaneechairat and Maleerat Maliyeam conducted the research and wrote the paper. All authors had approved the final version.

## REFERENCES


[1] T. Karoonboonyanan, *Standardization and Implementations of Thai Language*, Linguistics, 1999, pp. 1–11.
[2] Thai-language-history. (2019). [Online]. Available: https://www.todaytranslations.com/about/language-history/thai-language-history/
[3] Royal Institute, *Dictionary of the Royal Institute of Thailand*, Bangkok: Royal Academy, 2011.
[4] S. Sethaputra, *New Model English-Thai Dictionary*, Bangkok, 2005.
[5] W. Thiengburanathum, *NEW SE-ED's English-Thai & Thai-English Dictionary (Completely Revised & Updated Edition)*, Se-Education Public Company Ltd., 2008, pp. 1–1720.
[6] Top 5,000 word list. (2013). [Online]. Available: http://ling.arts.chula.ac.th/TNC/contents/File/freq-5000.xls
[7] J. Nilsen, *Thai Frequency Dictionary, The 4,000 Most Common Words in the Thai Language*, 2014.
[8] D. Sawamipuk, *Development of Thai Grammar Analysis Software under UNIX System*, Thammasat University Press, Bangkok, 1990.
[9] S. Raruenrom. "Dictionary-based Thai word separation," Senior Project Report, Dept. Computer Eng., Chulalongkorn University, Bangkok, 1991.
[10] V. Sornlertlamvanich, "Word segmentation for Thai in machine translation system," in *Machine Translation*, Bangkok: National Electronics and Computer Technology Center, 1993, pp. 50–56.
[11] A. Kawtrakul, C. Thumkanon, and S. Seriburi "A statistical approach to Thai word filtering," in *Proc. the 2nd Symposium on Natural Language Processing*, Bangkok, 1997, pp. 398–406.
[12] P. Chaloenpomsawat, "Feature-based Thai word segmentation," Master thesis, Chulalongkorn University, 1998.
[13] K. Thonglo, *Principles of Thai Language*, Bangkok: Ruamsan, 1994, pp. 214–291.
[14] T. Eckart and U. Quasthoff, "Statistical corpus and language comparison on comparable corpora," in *Building and Using Comparable Corpora*, S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung Eds. Berlin, Heidelberg: Springer, 2013, pp. 151–165.
[15] History of the Thai language. (1970). [Online]. Available: http://pennetpraphatsorn. Blogspot.com /2014/09/blog-post.html

**Hathairat Ketmaneechairat** received the PhD in electrical engineering from King Mongkut's University of Technology North Bangkok, Thailand. Currently, she is a lecturer at the College of Industrial Technology, King Mongkut's University of Technology North Bangkok. Her research areas are natural language processing and data mining, machine learning and artificial intelligence.

**Maleerat Maliyaem** received the PhD in information technology from King Mongkut's University of Technology North Bangkok, Thailand. Currently, she is a lecturer at the Faculty of Information Technology, King Mongkut's University of Technology North Bangkok. Her research areas are natural language processing and information retrieval, machine learning and artificial intelligence.